

Quarterly review

Research and Data

Q2 - 2015/16

Approximate team size during this quarter:

4.5 FTE, 2 research fellows, 10 collaborators

Time spent: strengthen 40%, focus 40%, experiment 20%

Q2 - Research and Data

Objective: Data ownership map



Objective	Measure of success	Status
FOCUS Streamline ownership of data and research <i>Team members involved: 1</i>	Publish a map for internal use of who is responsible for specific requests around research and analytics. (T112317)	completed

- Worked with individual teams on mentoring / backfilling / hiring of data analysts
- Published research/data ownership draft
- Socialized division of labor with Audience and Technology teams, Legal, Comms, C-team

https://office.wikimedia.org/wiki/Research_fingerpost

Research fingerpost

fin-ger-post
[noun] A board that shows the direction, and often distance, to a named place. (wiktionary)

Purpose of this document [edit | edit source]

The purpose of this document is to help you get as quickly as possible to the person or team who can support you with your data analysis and research needs. There is *no central team* at the Wikimedia Foundation responsible for all data and research-related questions. Information you may need when looking for data or research support may be found in one or more of many teams and departments, including:

- Product: *Community Tech, Discovery, Editing, Fundraising Tech, Reading*
- Technology: *Analytics, Design Research, Performance, Research and Data, Security*
- Advancement: *Online Fundraising*
- Community Engagement: *Learning and Evaluation*
- Legal
- Communications

We're creating this guide to simplify the process and help you find the most relevant contact person to direct your question to.

Still unsure? If you're still unsure, feel free to hop on the [#wikimedia--research](#) IRC channel or drop a line to research-internal@lists.wikimedia.org, our cross-departmental research list, where someone will be able to help.

This is a draft This document is constantly evolving, information may change rapidly as roles, processes and responsibilities progress.

Frequently Asked Questions [edit | edit source]



A hiking fingerpost in Corsica.

Q2 - Research and Data

Objective: Revscoring integration



Objective	Measure of success	Status
<p>STRENGTHEN</p> <p>Revscoring integration</p> <p><i>Team members involved: 1</i> <i>Collaborators: 9</i></p>	<p>Bring revscoring to fruition to our users as a Beta Feature (score integration into RC feed) (T112856)</p>	<p>missed</p> <p>Reason</p> <ul style="list-style-type: none">• blocked on code review for months• volunteer time went away during FR

- we weren't able to complete the design and deployment of this beta feature, due to limited resourcing and dependency on volunteers
- revscores integration into edit histories / RecentChanges feed pushed to Q3
- focus shifted on other priorities (massive deployments, impact analysis)
see **other accomplishments** below

https://meta.wikimedia.org/wiki/Research:Revision_scoring_as_a_service

Q2 - Research and Data

Revision scoring as a service

- Achieved up to **90-99% reduction in curation workload** [blog.wikimedia.de]
- **16 wikis currently supported**, incl. Wikidata, up from 6 in Q1 [meta.wikimedia.org]
- About **100 requests/minute** served [graphite.wikimedia.org]
- ORES announced via WMF blog [blog.wikimedia.org]
- Very large media attention [meta.wikimedia.org]
- Positive community adoption
- Study of disparate impacts of scores on anons



Q2 - Research and Data

Objective: Recsys productization

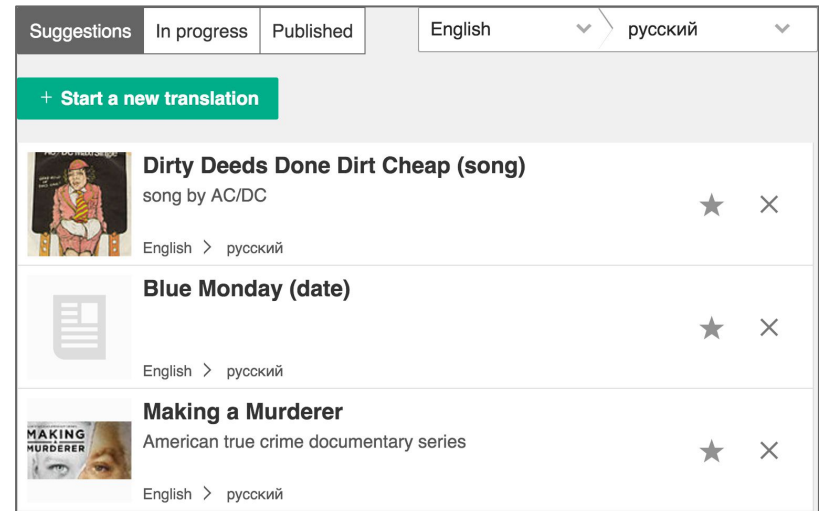


Objective	Measure of success	Status
<p><i>STRENGTHEN</i></p> <p>Productize article and link recommender systems</p> <p><i>Team members involved: 2</i> <i>Collaborators: 3</i></p>	<p>Article creation (T112321) and link (T112322) recommendations productized as services so they can be integrated with products and community tools.</p>	<p>completed</p>

- APIs available on Labs
- architecture and service implementation blessed by Ops
- First 3rd party tools adopting the service ([example](#))

Recommender systems integration

- Article recs **integrated into Content Translation** (December 2015)
- **16% of all saved translations** come from article recs
- Tested in small language set, API is now available for **all language pairs in Wikipedia**
- Integration of Link recommendations still in progress



Q2 - Research and Data

Objective: Value added research



Objective	Measure of success	Status
<i>EXPERIMENT</i> Value added research <i>Team members involved: 1</i>	Deliver research on historical data to quantify who adds value to English Wikipedia	missed Reason <ul style="list-style-type: none">completed part of the project (historical measurement of productivity on English WP)

- Finished historical analysis of productive edits by anons vs registered users (presented at [January 2016 Research Showcase](#))
- Partnership with external Hadoop as a service provider (Altiscale) ended in Q3 due to limited funding (currently exploring potential *pro bono* extension of services)

Q2 - Research and Data

Objective: Reader segmentation



Objective	Measure of success	Status
<p><i>EXPERIMENT</i></p> <p>Reader segmentation research STRETCH</p> <p><i>Team members involved: 3</i> <i>Collaborators: 1</i></p>	<p>Deliver research on reader segments and behavioral patterns (T112326)</p>	<p>completed</p>

- completed 3 contextual surveys and identified **robust categorization scheme** for reader motivation and information needs
- presented results to Reading team in preparation for Part 2 (Q3)

Q2 - Research and Data

Software engineering

Drafted template for **replicable productization** of research-oriented services [\[wikitech.wikimedia.org\]](https://wikitech.wikimedia.org)

Hired **Nathaniel Schaaf** as our first dedicated full-stack engineer.

Stage 0 (Architectural discussion)	
Horizontal scalability	allow additional load be taken up by simply adding new instances
Caching	Decide on caching and cache invalidation strategy
SPOF spotting / planning	Draw out general architecture, find SPOFs and think of ways to mitigate them
Stage 1 (Implementation)	
Actually build the code!	This is actual development, start building stuff!
Staging environment	provide an environment with the same set up as the Production environment, for test purposes
Deployment system	allow to deploy new changes confident that you can roll them back if they fail
Puppetized setup	allow spinning up new instances quickly
Comprehensive logging	identify bugs and errors more easily
Stage 2	
Metrics monitoring	define metrics that should go to graphite.wmflabs.org (examples of such metrics are number of revisions processed per minute, per wiki, % cached, etc.)
Scale hardening	Things to do to reduce amount of pages - for example, control the acceptance of new web requests when the celery queue is full. (This step is not required for all services).
Stage 3	
API documentation	More comprehensive document endpoints (API) and usage

Outreach

3 research papers and 1 poster accepted at major conferences in the field
CHI '16, CSCW '16, WSDM '16, WWW '16

3 research workshop proposals submitted and accepted
CSCW '16, ICWSM '16, WWW '16



Q2 - Research and Data

Category	Workflow	Comments	Type
	NDA's / endorsements	1 new MOU collaboration, 1 proposal reviewed for Legal	M
	Urgent DA requests	Supported Big English fundraising campaign in collaboration with Discovery	R
	Showcases and talks	2 research showcases, 1 brown bag, talks at WikiConference USA, University of Chicago, University of Missouri, Macalaster College, UMN	M

Type: new, reactive, maintenance

- [Research & Data team page](#)
 - Describing goals, processes and projects.
- [Goals for Q3 FY16](#)
 - What we are planning to do in the coming quarter
- [FY16 priorities](#)
 - Top priorities for the fiscal year
- [Phabricator workboard](#)
 - What we are currently doing